Stat 200 Friday March 2nd

The Bayesian Paradigm

The Bayesian Paradigm can be seen in some ways as an extra step in the modelling world just as parametric modelling is. We have seen how we could use probabilistic models to infer about some unknown aspect either by confidence intervals or by hypothesis testing.

The motivation for any statistical analyses is that some "target population" is not well understood-some aspects of it are unknown or unsure.

The idea in this paradigm is to say thta any uncertainty can be modelled in a probabilistic way.

It is true that there are very rarely situations when one doesn't know anything at all, asked to measure the table, you won't want to use a "pied de coulisse"(callipers) or a 100 yard measuring ribbon.

The probability model that we build can be quite approximate, it reflects one's beliefs and any prior experience we may have, it is described as personal or subjective.

Why ? Because it is different from person to person, examples that are easy to understand are about horse betting, the stock exchange...

So when the uncertainty about the model can be boiled down to a parameter θ the Bayesian statistician treats θ as if it were a random variable Θ whose distribution describes that uncertainty.

Elliciting a whole distribution may seem a challenge, in fact it's done by successive events of the type $\Theta \leq \theta$, and does NOT have to be very precise.

A subjective/personal probability is going to be subject to modification upon acquisition of further information supplied by experimanetal data.

Suppose a distribution with density $g(\theta)$ describes one's present uncertainties about some probability model with density $f(x|\theta)$.

Those uncertainties will change with the acquisition of data obtained by doing the experiment modelled by f.

Bayes theorem is essential in updating :

$$P(H|data) = \frac{P(data|H)P(H)}{P(data)}$$

The probability of H given the data is called the posterior probability of H, it is posterior to the data. The unconditional probability of $H : P(H)$ is the prior probability of H.

For given data P(data—H) is the likelihood of H.

For given data we often write :

$$P(H|data) \propto P(data|H)P(H)$$

The posterior is proportional to the likelihood time the prior.

If it helps (some people have a better understanding of odds):

$$\frac{P(H|data)}{P(H^c|data)} \propto \frac{P(data|H)}{P(data|H^c)} \frac{P(H)}{P(H^c)}$$

$\boxed{\text{Posterior odds = Prior odds times likelihood ratio.}}$ Now these formulae were written as if the rv were discrete for continuous random variables the behaviour is identical replacing probabilities with densities:

Represent the data by a random variable Y:

$$h(\theta\ ) \propto L(\theta\ )g(\theta\ )$$

$L(\theta\ )$ proportional to the probability density of Y given θ . In fact we can consider we are studying the joint distribution of two random variables Θ and Y.

The marginal distribution of Y is not exhibited, it is the proportionality factor. It can be written :

$$m(y) = \int f(y|tth)g(\theta\ )d\theta$$

<u>Remark</u>: One does NOT have to worry too much about the prior because as soon as the data comes in it is 'swamped' in the following sense: Two people with divergent prior opinions but reasonably open-minded will be forced into arbitrarily close agreement about future observations by a sufficient amount of data. We will see an example of this later on.

15.1 About Priors

"Gentle" priors reflect some agreed-upon weakness in the available information, for instance before any instruments went to the moon no one had any precise idea about the answer to the question: "How deep is the dust". The initial belief was overthrown as soon as any data came back.

When advance information is available the Bayesian method provides a routine way for updating uncertainty when new information comes in.

There are several steps to building a prior:

15.1.1 Calibrating degrees of belief

Suppose I wanted to discover "Your" probability that average adult male emperor penguins weigh more than 50 lbs? We will go through comparison experiments:

1. Would you rather bet on getting one green chip out of 1 R 1G or bet on A true?

   Suppose you prefer the latter.

2. Would you rather bet on getting a green chip out of 3G and 1 R ?

....etc... This allows for statements that enable us to bound probabilites.

Another type of thought experiment could be used to build $P[\Theta leq\theta]$ for an increasing sequence of θ's.

This is not usually how priors are built though because it seems quite an exhaustive process to build up a whole density prior, instead we are going to use families of priors who have easy updating processes with regards to the specific likelihoods at hand.

$$\boxed{\text{Posterior odds} = \text{Prior odds} \times \text{likelihood ratio.}}$$

$$h(\theta\ ) \propto L(\theta\ )g(\theta\ )$$

$L(\theta\ )$ proportional to the probability density of Y given θ . In fact we can consider we are studying the joint distribution of two random variables Θ and Y.

The marginal distribution of Y is not exhibited, it is the proportionality factor. It can be written :

$$m(y) = \int f(y|\theta\ )g(\theta\ )d\theta$$

15.2 Conjugate Priors

Sometimes a prior distribution can be approximated by one that is in a convenient family of distributions, which combines with the likelihood to produce a posterior that is manageable.

We see that an "objective" way of building priors for the binomial parameter was to use the 'conjugate family' distribution that has the property that the updated distribution is in the same family.

15.2.1 Binomial-Beta

15.3 Beta priors for the Binomial parameter

A little history: From Bayes 1763: A white billiard ball is rolled along a line and we look at where it stops, scale the table from 0 to 1. We suppose that it has a uniform probability of falling anywhere on the line. It stops at a point p.

A red billiard ball is then rolled n times under the same uniform assumption. r then denotes the number of times R goes less far than W went. Given X what inference can we make about p ?

Taken another way, we could have rolled n white balls first and then the red balls and looked hat how many balls there were before the red ball.

Or we could have rolled all the balls together and looked at where the red ball was, it is the a'th with probability 1/(n+1).

Let's say this again in our terminology: We are looking for the posterior distribution of p given X.

p is a number between 0 and 1

The prior distribution of p is Uniform(0,1)=Beta(1,1).

15.4 Beta family

$$f(p|r,s) \propto p^{r-1}(1-p)^{s-1}$$

$$B(r,s) = \int_0^1 p^{r-1}(1-p)^{s-1} = \frac{\Gamma(r)\Gamma(s)}{\Gamma(r+s)}$$

$X \sim \mathcal{B}\ (n,p)$

$$P(X=x|p) = \binom{n}{x} p^x(1-p)^{n-x}$$

$$P(a < p < b and X = x) = \int_a^b \binom{n}{x} p^x(1-p)^{n-x} dp$$

$$P(X=x) = \int_0^1 \binom{n}{x} p^x(1-p)^{n-x} dp$$

$$P(a < p < b|X=x) = \frac{\int_a^b \binom{n}{x} p^x(1-p)^{n-x} dp}{B(x+1, n-x+1)}$$

15.5 Normal-Normal

Some of you may have seen in section but I will mention it here as it is very important: The conjugate for a Normal likelihood is the Normal distribution; here is the theorem, I will not prove it, its simple algebra, and its in the book page 590.

For practical reasons, we define the precision as the inverse of the variance: we denote by $\xi = \frac{1}{\sigma^2}$ and $\xi_0 = \frac{1}{\sigma_0^2}$

Theorem 15.1 *Suppose that* $\mu \sim \mathcal{N}\ (\mu_0, \sigma_0^2)$. *Then the posterior distribution of* μ *is normal with mean*

$$\mu_1 = \frac{\xi_0\mu_0 + \xi x}{\xi_0 + \xi}$$

and precision

$$\xi_1 = \xi_0 + \xi$$

The posterior mean is a weighted average of the prior mean and the data, weights being proportional to the respective precisions.

With a very gentle prior we would have a very low precision ξ_0, a very flat prior and mostly the posterior is Normal with x as its mean.

Of course what we are usually interested in is the posterior given an iid sample of size n, what you could expect happens it is equivalent to adding one observation $\bar{x}$ from a distribution that has variance σ^2/n.

15.6 Multinomial-Dirichlet

You are given a set $\mathcal{X}$ (here taken as finite) and a probability density $p(x), (p(x) \geq 0, \sum p(x) = 1)$. Also given is a set A in $\mathcal{X}$. The problem is to compute or approximate $p(A)$. We consider one of the three basic problems used throughout by Feller - the Birthday Problem, the Coupon Collector's problem and the Matching Problem. This will be considered from a Bayesian standpoint.

In order to go further we need to extend what we did before for the binomial and its Conjugate Prior to the multinomial and the the Dirichlet Prior. This is a probability distribution on the n simplex

$$\Delta_n = \{\tilde{p} = (p_1, \cdots, p_n),\ p_1 + \cdots + p_n = 1,\ p_i \geq 0\ \}$$

It is a n-dimensional version of the beta density. The Dirichlet has a parameter vector: $\tilde{a} = (a_1, \ldots, a_n)$. Throughout we write $A = a_1 + \cdots + a_n$.

Δ_n is normalised to have total mass 1 the Dirichlet has density:

$$D_{\tilde{a}}(\tilde{x}) = \frac{\Gamma(A)}{\prod \Gamma(a_i)} x_1^{a_1-1} x_2^{a_2-1} \cdots x_n^{a_n-1}$$

The uniform distribution on Δ_n results from choosing all $a_i = 1$. The multinomial distribution corresponding to k balls dropped into n boxes with fixed probability $(p_1, \cdots, p_n)$ (with the ith box containing k_i balls) is

$$\binom{k}{k_1 \ldots k_n} p_1^{k_1} \cdots p_n^{k_n}$$

If this is averaged with respect to $\mathbf{D}_{\tilde{a}}$ one gets the marginal (or Dirichlet/ Multinomial):

$$P(k_1, \ldots, k_n) = P((k_1, k_2, \ldots, k_n)) = \frac{(a_1)_{(k_1)}(a_2)_{(k_2)} \ldots (a_n)_{(k_n)}}{A_{(k)}}$$

$$\text{where} \quad m_{(j)} \overset{\text{def}}{=} m(m+1)\cdots(m+(j-1))$$

From a more practical point of view there are two simple procedures worth recalling here:

- To pick $\tilde{p}$ from a Dirichlet prior; just pick $X_1, X_2, \ldots, X_n$ independant from gamma densities

$$\frac{e^{-x}x^{a_i-1}}{\Gamma(a_i)} \text{ and set } p_i = \frac{X_i}{X_1 + \cdots X_n}, 1 \le i \le N$$

- To generate sequential samples from the marginal distribution use **Polya's Urn**: Consider an urn containing a_i balls of color i (actually fractions are allowed).

  Each time, choose a color i with probability proportional to the number of balls of that color in the urn. If i is drawn, replace it along with another ball of the same color.

The Dirichlet is a convenient prior because the posterior for $\tilde{p}$ having observed $(k_1, \cdots, k_n)$ is Dirichlet with probability $(a_1+k_1, \cdots, a_n+k_n)$. An important characterization of the Dirichlet: it is the only prior that predicts outcomes linearly in the past. One frequently used speical case is the symmetric Dirichlet when all $a_i = c > 0$. We denote this prior as $\mathbf{D}_c$.